Estimating the benefits of student model improvements on a substantive scale

Michael V. Yudelson
Carnegie Learning, Inc.
437 Grant St.
Pittsburgh, PA 15219, USA
myudelson@carnegielearning.com

Kenneth R. Koedinger
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA 15213, USA
koedinger@cmu.edu

ABSTRACT

Educational Data Mining researchers use various prediction metrics for model selection. Often the improvements one model makes over another, while statistically reliable, seem small. The field has been lacking a metric that informs us on how much practical impact a model improvement may have on student learning efficiency and outcomes. We propose a metric that indicates how much wasted practice can be avoided (increasing efficiency) and extra practice would be added (increasing outcomes) by using a more accurate model. We show that learning can be improved by 15-22% when using machine-discovered skill model improvements across four datasets and by 7-11% by adding individual student estimates to Bayesian Knowledge Tracing.

1. INTRODUCTION

In this work we are discussing an approach that translates differences in statistical metrics between the two models into the potential differences in the number of practice attempts students would be prescribed and the time students could allocate more optimally if a better-fitting student model is deployed. We consider two types of model comparisons. First, we compare alternative skill models of the problem domain while keeping the student modeling algorithm the same. Second, we compare keeping the skill model the same. Second, keeping the skill model the same and changing the student modeling algorithm. We discuss results obtained for several datasets that cover domains such as middle school algebra and geometry, English, and numberline games. Our investigation shows that, despite the improvement in model accuracy metric being seemingly small, representing the differences in terms of missed practice opportunities and time reveals substantial differences.

2. DATA

We used the datasets from the KDD Cup 2010 EDM Challenge and from the Pittsburgh Science of Learning Center (PSLC) DataShop (`www.pslcdatashop.org`): Algebra I dataset, and Bridge to Algebra collected in 2008-09. We also used 4 PSLC DataShop datasets addressing Geometry (1996-97 and 2010), Articles (2009), and Numbeline Games (2011). The KDD Cup 2010 data was donated by Carnegie Learning Inc. PSLC DataShop datasets were collected by various researcher partners of PSLC (`www.learnlab.org`). The KDD Cup 2010 Algebra I dataset has 8,918,054 practice attempts of 3,310 students and has 2 skill models: 'KTraced-Skills' (kts) used in cognitive tutor, and an alternative 'Sub-Skills' (ss) model. The KDD Cup 2010 Bridge to Algebra dataset contains data of 6,043 students comprised of 20,012,498 rows and has the two skill models as well. PSLC DataStop Geometry 1996-97 data covers of 5,388/59 transactions/students, Articles 2009 data - 6,887/120, Geometry 2010 - 140,854/120, Numberline games 2011 data - 4,341/51.

3. MODELS

We will use Bayesian Knowledge Tracing (BKT) [2] to fit models of student learning. BKT is a method often used in Intelligent Tutoring Systems (ITS). In addition to standard BKT, we will use an individualized BKT (iBKT) models described in [3]. Namely, the model where the *p-learn* has a per-skill and per-student component. We implemented a tool capable of fitting standard and individualized BKT models on large datasets (such as KDD Cuo 2010 dataset) in an efficient way. Our tool is implemented in C/C++ and can fit BKT models from large datasets very quickly. For more details please refer to [3]).

4. METHOD

First, we fit original and alternative models for all of the datasets and skill modelswe have. For the two KDD Cup 2010 datasets we fit BKT and iBKT models. For the four DataShop datasets, we fit BKT model only. Out of several skill models available for each DataShop dataset we select original one and the best skill model discovered using a human-machine Learning Factors Analysis procedure [1].

We then compute probabilities of skill mastery for all student attempts. We used a threshold probability of 0.95 (a traditionally accepted value) to determine the moment of mastery. If, according to the model, student did not reach mastery for a particular skill within the recorded student data, we calculate the number of under-practice attempts. If student's skill reaches mastery earlier than the latest attempt recorded, we compute the number of over-practice attempts. The mastery data is aggregated by student taking under-practice attempts into consideration.

Table 1: Comparing models in terms of root mean squared error, percent cases number of prescribed practice opportunities differs by at least one, average student/skill practice opportunities, and time

(a) Estimated prediction improvements and *practical benefits* of replacing hand-made by LFA machine-discovered KC models across four DataShop datasets (RMSE values are given for a student-stratified 10-fold cross-validation).

	Time	KCs		RMSE		% diff Orig.-LFA			Mean stud. opp/KC				Stud. time		
Dataset	/step	Orig.	LFA	Orig.	LFA	≤ -1	(-1,1)	≥ 1	Orig.	LFA	diff	%	total	diff	%
Geometry 1996-97	17.12s	15	18	0.410	0.400	10%	29%	61%	8.8	7.8	1.5	18-20	26m	2m	9
Articles 2009	15.09s	13	26	0.437	0.420	5%	53%	43%	7.9	7.1	1.2	15-17	14m	3m	23†
Geometry 2010	15.10s	46	43	0.240	0.239	95%	5%	0%	8.6	10.5	1.9	18-22	88m	6m	7
Numberline 2011	12.77s	12	22	0.459	0.457	41%	22%	37%	15.1	15.2	2.8	19	18m	32m	182†

†These values could be inflated due to absence of mastery learning in respective tutors and as a result the amount of student work being less optimal.

(b) Estimated prediction improvements and *practical benefits* of replacing standard BKT models by individualized BKT models across two KDD Cup 2010 datasets (RMSE values are given for a student-stratified 10-fold cross-validation).

	Time		RMSE		% diff BKT-iBKT			Mean stud. opp/KC				Stud. time		
Dataset	/step	KCs	BKT	iBKT	≤ -1	(-1,1)	≥ 1	BKT	iBKT	diff	%	total	diff	%
Algebra 1(kts)	n/a	515	0.363	0.361	24%	72%	4%	12.1	12.9	1.1	9	n/a	n/a	n/a
Algebra 1(ss)	n/a	541	0.342	0.341	34%	63%	3%	12.5	13.6	1.4	10-11	n/a	n/a	n/a
B.to Algebra(kts)	12.81s	807	0.363	0.359	22%	74%	5%	14.3	14.9	1.0	7	361m	15m	4
B.to Algebra(ss)	12.81s	933	0.359	0.355	27%	68%	5%	18.3	19.2	1.2	7	485m†	22m	5

†Difference in times between *its* and *ss* KC models is due to change in the subset of data selected.

Finally, we compute the time it takes a student to solve one tutor step. This time is used to compute the typical length of all student sessions in the system. Having the sum of the number of practice opportunities it takes the student to master all skills (correcting for under-practice) from the both models being compared and plugging in the average step duration, we compute the overall amount of time student *wastes* for under-practicing and over-practicing.

5. RESULTS AND DISCUSSION

Table 1 is a summary of model comparisons. Table 1a compares original skill models and best fitting machine-discovered skill models for the DataShop datasets. Table 1b compares standard BKT and individualized BKT modeling methods for the same skill models in KDD Cup 2010 datasets. Despite the vast difference in the size of the datasets (inherently the size of curriculum), improvements with respect to student-stratified cross-validated RMSE are quite small. Just like the improvements in RMSE, the mean absolute difference in mean student opportunities are small: from 1.1 to 2.8 practice attempts. However, in terms of percent practice opportunities, those differences constitute 15-22% in DataShop datasets, and 7-11% in KDD Cup 2010 datasets.

The practice opportunity differences are shown in Table 1a and Table 1b under *% diff Orig-LFA* and *% diff BKT-iBKT* respectively. Here the column marked '≤ -1' indicates the percent of student-KC experiences for which the model built on the LFA-discovered KC model prescribes at least one opportunity /less/ on average than the model built on the Original KC model. Similarly, column '≥ 1' indicates the percent that the LFA-discovered skill model prescribes at least one *more* opportunity..

The overall amount of time students spend with the tutor differs from dataset to dataset: from 14-18 mimutes to 8 hours. The absolute and percent time values for time differences in Table 1 reflect both over-practice and under-practice together. The absolute average and percent average time difference between the models are given next to the total time students spend o average. The percent of the time students are *wasting* is 7-9% on Geometry DahaShop datasets (able 1a) and 4-5% on Bridge to Algebra KDD Cup 2010 dataset (Table 1b). A higher values of nearly a quarter of time misused (23%) in the case of Articles 2009 dataset and almost twice the time (182%) misused on the Numberline 2011 DataShop dataset, are due to the fact that both did not implement mastery learning and, contrary to the cases of tutors used to collect other datasets, problems were not sequenced in attempt to maximize students' learning.

6. ACKNOWLEDGMENTS

This work was supported by the PSLC DataShop team, Carnegie Learning Inc., and National Science Foundation (award #SBE-0836012). Most of the work was done when the first author was at Carnegie Mellon University.

7. REFERENCES

[1] H. Cen, K. R. Koedinger, and B. Junker. Learning factors analysis - a general method for cognitive model evaluation and improvement. In M. Ikeda, K. D. Ashley, and T.-W. Chan (Eds.), *8th International Conference on Intelligent Tutoring Systems (ITS 2006)*, pp. 164–175, Jhongli, Taiwan, June 26-30, 2006. Springer.

[2] A. T. Corbett and J. R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4):253–278, 1995.

[3] M. V. Yudelson, K. R. Koedinger, and G. J. Gordon. Individualized bayesian knowledge tracing models (in press), In: Artificial Intelligence in Education, 2013.